Alternating Direction Method of Multipliers for distributed subgroup analysis

Ying Lin

2024-03-14

This post briefly introduces how to apply Alternating Direction Method of Multipliers (ADMM) to solve the distributed subgroup analysis, particularly the one regularized by fused lasso.

1 Introduction

Given a communication network (a bidirected graph) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$[1] with the number of nodes $|\mathcal{V}| = K$, consider the following distributed optimization problem

$$\min_{\mathbf{x}:=\{x_k \in \mathbb{R}^d\}_{k\in\mathcal{V}}} \sum_{k\in\mathcal{V}} f_k(x_k) + \lambda \sum_{(i,j)\in\mathcal{E}} \|x_i - x_j\|_1, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{dK}$ is the global target variables with x_k being the local model parameter for node $k \in \mathcal{V}$; $f_k : \mathbb{R}^d \to \mathbb{R}$ is the convex but not necessary smooth local loss function for node $k \in \mathcal{V}$; $\|\cdot\|_1$ is the 1-norm. Examples of f_k include *least squares* and *hinge loss function* (for training *support vector machine*). In this post, we consider using Alternating Direction Method of Multipliers (ADMM) to solve this problem in the distributed setting without any restriction on communication.

In the distributed setting, a node $k \in \mathcal{V}$ can only access its local information (e.g., local loss function f_k and local variables x_k), and obtain limited information via communication with its neighbors (e.g., x_j for $j \in \mathcal{N}_k$ where $\mathcal{N}_k$ is the set of nodes adjacent to node k).

We first deduce the dual problem of (1). Then the ADMM will be developed.

[1] Note that $(i, j) \in \mathcal{E}$ if and only if $(j, i) \in \mathcal{E}$. By convention, we assume that $(i, i) \notin \mathcal{E}$ for all $i \in \mathcal{V}$.

2 Dual problem and ADMM

Proposition 2.1. *The dual problem of (1) is*

$$\begin{aligned} \max_{\mathbf{w}:=\{w_{ij}\}_{(i,j)\in\mathcal{E}}} \quad & -\sum_{k\in\mathcal{V}} f_k^*\Big(\sum_{j\in\mathcal{N}_k}(w_{jk}-w_{kj})\Big) \\ s.t. \quad & \|w_{ij}\|_\infty \le \lambda \quad \forall (i,j)\in\mathcal{E}, \end{aligned} \tag{2}$$

where $\mathbf{w} := \{w_{ij}\}_{(i,j)\in\mathcal{E}} \in \mathbb{R}^{d|\mathcal{E}|}$ *contains the dual variables;* f_k^* *is the convex conjugate of* f_k*;* $\mathcal{N}_k$ *is the set of nodes adjacent to node* $k \in \mathcal{V}$*;* $\|\cdot\|_\infty$ *is the infinity norm.*

Proof. We first rewrite (1) as the following equivalent constraint optimization problem

$$\begin{aligned} \min_{\mathbf{x},\mathbf{d}} \quad & \sum_{k\in\mathcal{V}} f_k(x_k) + \lambda \sum_{(i,j)\in\mathcal{E}} \|d_{ij}\|_1 \\ \text{s.t.} \quad & d_{ij} = x_i - x_j \quad \forall (i,j)\in\mathcal{E}, \end{aligned} \tag{3}$$

where $\mathbf{d} := \{d_{ij}\}_{(i,j)\in\mathcal{E}} \in \mathbb{R}^{d|\mathcal{E}|}$ is the dummy variable. The Lagrangian function of (3) is

$$\mathcal{L}(\mathbf{x},\mathbf{w}) = \sum_{k\in\mathcal{V}} f_k(x_k) + \lambda \sum_{(i,j)\in\mathcal{E}} \|d_{ij}\|_1 + \sum_{(i,j)\in\mathcal{E}} \langle w_{ij}, x_i - x_j - d_{ij}\rangle,$$

where $\mathbf{w} := \{w_{ij}\}_{(i,j)\in\mathcal{E}} \in \mathbb{R}^{d|\mathcal{E}|}$ is the dual variable.

Notice that we have

$$\min_{x_k} \quad f_k(x_k) - \langle x_k, \sum_{j\in\mathcal{N}_k}(w_{jk}-w_{kj})\rangle = -f_k^*\Big(\sum_{j\in\mathcal{N}_k}(w_{jk}-w_{kj})\Big);$$

$$\min_{d_{ij}} \quad \lambda\|d_{ij}\|_1 - \langle w_{ij}, d_{ij}\rangle = \begin{cases} 0 & \text{if } \|w_{ij}\|_\infty \le \lambda, \\ -\infty & \text{otherwise.} \end{cases}$$

Therefore, we have the dual problem as in (2). □

Next we move on to the development of the corresponding ADMM.

Given the augmented parameter $\beta > 0$, consider the augmented Lagrangian function with respect to (3), defined as

$$\begin{aligned} \mathcal{L}_\beta(\mathbf{x},\mathbf{d},\mathbf{w}) := \sum_{k\in\mathcal{V}} f_k(x_k) + \lambda \sum_{(i,j)\in\mathcal{E}} \|d_{ij}\|_1 + \sum_{(i,j)\in\mathcal{E}} \langle w_{ij}, x_i - x_j - d_{ij}\rangle \\ + \frac{\beta}{2} \sum_{(i,j)\in\mathcal{E}} \|x_i - x_j - d_{ij}\|^2. \end{aligned}$$

For simplicity, we make use of the *incidence matrix* $M \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ and the *Laplacian matrix* $L = M^\top M \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ of the communiaction graph $\mathcal{G}$, which are defined by

$$M_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathcal{E}, \\ -1 & \text{if } (j,i) \in \mathcal{E}, \\ 0 & \text{otherwise;} \end{cases} \qquad L_{ij} = \begin{cases} |\mathcal{N}_i| & \text{if } i = j, \\ -1 & \text{if } (i,j) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases}$$

Using M, we can simplify the augmented Lagrangian function as

$$\begin{aligned} \mathcal{L}_\beta(\mathbf{x},\mathbf{d},\mathbf{w}) =& \sum_{k\in\mathcal{V}}(f_k(x_k) + \langle x_k, \sum_{j\in\mathcal{N}_k}(w_{kj} - w_{jk})\rangle) \\ &+ \sum_{(i,j)\in\mathcal{E}}(\lambda\|d_{ij}\|_1 - \langle w_{ij}, d_{ij}\rangle + \frac{\beta}{2}\|x_i - x_j - d_{ij}\|^2) \\ =& \sum_{k\in\mathcal{V}}(f_k(x_k) + \langle x_k, \sum_{j\in\mathcal{N}_k}(w_{kj} - w_{jk})\rangle) \\ &+ \lambda\|\mathbf{d}\|_1 - \langle \mathbf{w}, \mathbf{d}\rangle + \frac{\beta}{2}\|(M\otimes I_d)\mathbf{x} - \mathbf{d}\|^2 \\ =& \sum_{k\in\mathcal{V}}(f_k(x_k) + \langle x_k, \sum_{j\in\mathcal{N}_k}(w_{kj} - w_{jk})\rangle) \\ &+ \lambda\|\mathbf{d}\|_1 - \langle \mathbf{w}, \mathbf{d}\rangle + \frac{\beta}{2}\|(M\otimes I_d)\mathbf{x}\|^2 - \beta\langle (M\otimes I_d)\mathbf{x}, \mathbf{d}\rangle + \frac{\beta}{2}\|\mathbf{d}\|^2, \end{aligned}$$

where I_d is the $d \times d$ identity matrix.

The classical ADMM applied to (3) consists of the update formulae

$$\begin{cases} \mathbf{x}^{t+1} = \operatorname{argmin}_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{d}^t, \mathbf{w}^t), \\ \mathbf{d}^{t+1} = \operatorname{argmin}_{\mathbf{d}} \mathcal{L}_\beta(\mathbf{x}^{t+1}, \mathbf{d}, \mathbf{w}^t), \\ \mathbf{w}^{t+1} = \mathbf{w}^t + \beta((M\otimes I_d)\mathbf{x}^{t+1} - \mathbf{d}^{t+1}). \end{cases}$$

One can readily see that the $\mathbf{x}$-update admits the form

$$\mathbf{x}^{t+1} = \operatorname*{argmin}_{\mathbf{x}} \sum_{k\in\mathcal{V}}(f_k(x_k) + \langle x_k, \sum_{j\in\mathcal{N}_k}(w_{kj}^t - w_{jk}^t)\rangle) + \frac{\beta}{2}\|(M\otimes I_d)\mathbf{x}\|^2 - \beta\langle (M\otimes I_d)\mathbf{x}, \mathbf{d}^t\rangle,$$

where $(M\otimes I_d)\mathbf{x}$ involves the full knowledge of the graph $\mathcal{G}$. However, the full information of the graph $\mathcal{G}$ is unrevealed to any of nodes in the distributed setting. It follows that in this case the classical ADMM possesses more than three blocks, which does not converge in general. To address this issue, we use the proximal ADMM by adding a proximal term in the $\mathbf{x}$-update subproblem to make the subproblem solvable in parallel (Fazel et al. 2013; Li and Pong 2015). For this approach, we require a valid Bregman function, which we now define.

Let

$$\phi(\mathbf{x}) = \beta\|(N \otimes I_d)\mathbf{x}\|^2 - \frac{\beta}{2}\|(M \otimes I_d)\mathbf{x}\|^2,$$

where $N \in \mathbb{R}^{K\times K}$ is the diagonal matrix whose diagonal elements are $N_{ii} = \sqrt{|\mathcal{N}_i|}$. The Hessian of ϕ is hence given by

$$\nabla^2\phi(\mathbf{x}) = 2\beta(N^\top N \otimes I_d) - \beta(M^\top M \otimes I_d) = (2\beta N^\top N - \beta M^\top M) \otimes I_d =: P \otimes I_d.$$

Recall that $M^\top M = L$ and notice that $2\beta N^\top N$ is again a diagonal matrix whose i-th diagonal element is $2\beta|\mathcal{N}_i|$. We can then see that

$$P_{ij} = \begin{cases} \beta|\mathcal{N}_i| & \text{if } i = j \\ \beta & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}.$$

Hence it holds that $\nabla^2\phi(\mathbf{x}) \succeq 0$ thanks to the Gershgorin circle theorem. Thus, ϕ is a valid kernel for defining the Bregman distance

$$D_\phi(\mathbf{x}_1, \mathbf{x}_2) = \phi(\mathbf{x}_1) - \phi(\mathbf{x}_2) - \langle\nabla\phi(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2\rangle.$$

Now, the proximal ADMM applied to (3) is

$$\begin{cases} \mathbf{x}^{t+1} = \operatorname{argmin}_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{d}^t, \mathbf{w}^t) + D_\phi(\mathbf{x}, \mathbf{x}^t), \\ \mathbf{d}^{t+1} = \operatorname{argmin}_{\mathbf{d}} \mathcal{L}_\beta(\mathbf{x}^{t+1}, \mathbf{d}, \mathbf{w}^t), \\ \mathbf{w}^{t+1} = \mathbf{w}^t + \beta((M \otimes I_d)\mathbf{x}^{t+1} - \mathbf{d}^{t+1}). \end{cases}$$

For the $\mathbf{x}$-update, we can see that the corresponding minimization problem now decouples as

$$\begin{aligned}
&\mathbf{x}^{t+1} = \operatorname*{argmin}_{\mathbf{x}} \mathcal{L}_\beta(\mathbf{x}, \mathbf{d}^t, \mathbf{w}^t) + D_\phi(\mathbf{x}, \mathbf{x}^t) \\
=\ &\operatorname*{argmin}_{\mathbf{x}} \sum_{k\in\mathcal{V}} \left\{ f_k(x_k) + \langle x_k, \sum_{j\in\mathcal{N}_k} (w_{kj}^t - w_{jk}^t)\rangle \right\} + \frac{\beta}{2}\|(M \otimes I_d)\mathbf{x}\|^2 - \beta\langle(M \otimes I_d)\mathbf{x}, \mathbf{d}^t\rangle \\
&\qquad + \beta\|(N \otimes I_d)\mathbf{x}\|^2 - \frac{\beta}{2}\|(M \otimes I_d)\mathbf{x}\|^2 - \langle(P \otimes I_d)\mathbf{x}^t, \mathbf{x}\rangle \\
=\ &\operatorname*{argmin}_{\mathbf{x}} \sum_{k\in\mathcal{V}} \left\{ f_k(x_k) + \langle x_k, \sum_{j\in\mathcal{N}_k} (w_{kj}^t - w_{jk}^t - \beta d_{kj}^t + \beta d_{jk}^t - \beta x_j^t) - \beta|\mathcal{N}_k|x_k^t\rangle + \beta|\mathcal{N}_k|\|x_k\|^2 \right\} \\
=\ &\left(\operatorname*{argmin}_{x_k} \left\{ f_k(x_k) + \beta|\mathcal{N}_k| \cdot \left\| x_k - \frac{\beta|\mathcal{N}_k|x_k^t - \sum_{j\in\mathcal{N}_k}(w_{kj}^t - w_{jk}^t - \beta d_{kj}^t + \beta d_{jk}^t - \beta x_j^t)}{2\beta|\mathcal{N}_k|} \right\|^2 \right\} \right)_{k\in\mathcal{V}}
\end{aligned}$$

$$= \left(\text{prox}_{\frac{f_k}{2\beta|\mathcal{N}_k|}} \left(\frac{\beta|\mathcal{N}_k| x_k^t - \sum_{j\in\mathcal{N}_k} (w_{kj}^t - w_{jk}^t - \beta d_{kj}^t + \beta d_{jk}^t - \beta x_j^t)}{2\beta|\mathcal{N}_k|} \right) \right)_{k\in\mathcal{V}}.$$

The $\mathbf{d}$-update remains unchanged, and is given by

$$\begin{aligned}
&\mathbf{d}^{t+1} = \operatorname*{argmin}_{\mathbf{d}} \mathcal{L}_\beta(\mathbf{x}^{t+1}, \mathbf{d}, \mathbf{w}^t) \\
&= \operatorname*{argmin}_{\mathbf{d}} \left\{ \sum_{(i,j)\in\mathcal{E}} \left(\lambda \|d_{ij}\|_1 - \langle d_{ij}, w_{ij}^t + \beta(x_i^t - x_j^t) \rangle + \frac{\beta}{2} \|d_{ij}\|^2 \right) \right\} \\
&= \left(\text{prox}_{\frac{\lambda}{\beta}\|\cdot\|_1} \left(\frac{w_{ij}^t + \beta(x_i^t - x_j^t)}{\beta} \right) \right)_{(i,j)\in\mathcal{E}}.
\end{aligned}$$

Therefore, we obtain the ADMM to solve (1).

Reference

Fazel, Maryam, Ting Kei Pong, Defeng Sun, and Paul Tseng. 2013. "Hankel Matrix Rank Minimization with Applications to System Identification and Realization." *SIAM Journal on Matrix Analysis and Applications* 34 (3): 946–77. https://doi.org/10.1137/110853996.

Li, Guoyin, and Ting Kei Pong. 2015. "Global Convergence of Splitting Methods for Nonconvex Composite Optimization." *SIAM Journal on Optimization* 25 (4): 2434–60. https://doi.org/10.1137/140998135.